Robust MDPs with Applications in Fisheries Management

Anna Poulton & Jeet Dhoriyani

March 2022

Contents

1 Introduction

Markov decision processes (MDPs) are an important tool for dynamic decision making. However, they may produce inaccurate results if some model parameters (e.g., the transition probabilities) are not calculated precisely; in real world scenarios there are often sampling errors and model misfits which could make MDPs produce far from optimal policies [6, 16]. Robust Markov decision processes (RMPDs) provide an important framework for overcoming this uncertainty. In the robust framework, the parameters for an MDP are considered to come from an uncertainty set, from which they can take any plausible value. The decision maker is assumed to be risk-averse to this uncertainty, and wants to make decisions that maximize some goal (e.g., the expected discounted reward) in the worst-case scenario [16].

There are many possible applications of MDPs in the real world. For example, MDPs are widely applicable in financial trading [1]. The agent might be a person (or an algorithm) who considers the price of an asset as the state, and takes actions to buy or sell the asset. Potential profit from the trade is considered as the reward, and the transition probabilities could be calculated on the basis of the historic data and market sensitivity. It may reasonable to assume that the agent is 'risk-averse' to uncertainty in the problem (in asset pricing, market volatility, etc.) and would prefer to 'play it safe' by maximizing the profit in the worst-case scenario. In this case, the robust MDP framework would be a reasonable modeling choice [16].

Other important applications of MDPs range from healthcare to farming to fisheries management [14], the latter of which will be our focus. While most modeling work for fisheries using MDPs takes a nonrobust approach, there are several reasons why a robust approach would be reasonable and provide valuable management advice in a fisheries context. For one thing, there is much uncertainty involved when estimating characteristics of a fish population (e.g., birth/death/growth rates) and of the corresponding fishing fleet (e.g., how a certain fishing intensity may impact the fish population). Furthermore, risk aversion is a reasonable response to this uncertainty for both fishers and fishery managers, where it is very desirable to avoid worst-case fishing outcomes (heavy depletion or extinction, with near total loss of profitability).

The remainder of the paper is outlined as follows. In Section 2, we give a mathematical description of the MDP and robust MDP frameworks. In Section 3, we describe the simplest category of uncertainty sets, which use the assumption of (s,a)-rectangularity. More categories of uncertainty sets are described in Section 4, and further extensions are considered in Section 5. In Section 6 we describe and solve a fisheries MDP in robust and nonrobust settings, and finally our work is concluded in Section 7.

2 MDPs and Robust MDPs

A Markov decision process (MDP) involves a decision maker taking actions to affect the rewards and transition probabilities of a Markov chain as it moves between states. Here we give a mathematical description of MDPs in nonrobust and robust settings. For consistency, we will mostly use the notation in Wiesemann et al. 2013. We will be primarily focused on MDPs with finite state and action spaces: for simplicity, we let $\mathcal{S} = \{1, \dots, S\}$ denote the state space and $\mathcal{A} = \{1, \ldots, A\}$ denote the action space. The transition probabilities from each state-action pair will be denoted $P_{s,a} = p(.|s, a)$, with the overall transition kernel denoted as P. The reward given after transitioning from state s_t to s_{t+1} and taking action a_t is represented by $r(s_t, a_t, s_{t+1})$. We will also mostly focus on the infinite horizon discounted setting (with discount factor λ), although we also provide some results pertaining to the finite horizon setting.

The typical goal of an MDP is to find an optimal policy for the decision maker. The most broad definition of a policy is a mapping from the history of the MDP to a probability distribution of actions, but often policies fall into simpler classes: Markovian policies (the decision only depends on the current state s_t and not prior history), stationary policies (a Markovian policy where the decision is not time-dependent), and deterministic policies (the decision is a single action rather than a distribution of actions) [6]. Combining the transition probabilities with a selected policy π induces a Markov chain.

The performance of a policy π can be quantified by considering the value function (or 'reward-to-go' function), which gives the expected total discounted reward as a function of the starting state s under the transitions P [16].

$$v_s(\pi; P) = \mathbb{E}^{\pi,P}\left[\sum_{t=0}^{\infty} \lambda^t r(s_t, a_t, s_{t+1}) | s_0 = s\right] \tag{1}$$

In the nonrobust setting, the decision maker seeks a policy π^* that maximizes $v_s(\pi; P)$. (Note that we could also specify an initial distribution over states, p_0, rather than a single starting state.)

In many applications only noisy, inexact data may be available to estimate model parameters, and the nonrobust policy generated under these estimated parameters could perform badly in the true setting [6, 16]. Past works have looked at MDPs under specific distributions of uncertainty, but it may be hard to predict these distributions ahead of time [12]. Modeling the transition matrix as a polytope is another option, but this method may not be practical due to tractability and computational limits [5, 15].

This motivates our robust approach, which differs from the nonrobust approach in two key ways. First, it is assumed that the transition probabilities are uncertain, and can take any value in some 'uncertainty set' (or 'ambiguity set') [16]. We will use $\mathscr{P}$ to represent the uncertainty set on the transition kernel. We will generally assume that the uncertainty set $\mathscr{P}$ is compact and convex, although we later discuss results if this is not the case. Second, the decision maker is assumed to be risk-averse to this uncertainty, and seeks a policy that maximizes the discounted reward in the worst case scenario [16]. More formally, they seek a π^* such that

$$R(\pi^*) = \sup_{\pi} R(\pi), \quad \text{where} \quad R(\pi) = \inf_{P \in \mathscr{P}} \mathbb{E}^{\pi,P}\left[\sum_{t=0}^{\infty} \lambda^t r(s_t, a_t, s_{t+1}) | s_0 \sim p_0\right] \tag{2}$$

Finding a policy to maximize $R(\pi)$ is a key problem called the 'robust policy improvement problem'; we will also discuss the inner problem, of how to evaluate $R(\pi)$ for a specific π, called the 'robust policy evaluation problem' [16]. Note that relating Equations 1 and 2, we have $R(\pi) = \inf_{P \in \mathscr{P}}\{p_0^T v_s(\pi; P)\}$ [16].

One of the simplest possible assumptions for uncertainty sets, '(s,a)-rectangularity', will be considered in the next section. We will discuss key results for this class of uncertainty sets, along with several examples. In Section 4, we will discuss a few tractable generalizations of (s,a)-rectangularity, along with some results for the general case of uncertainty sets.

3 (s,a)-rectangularity

The formal definition of an (s,a)-rectangular uncertainty set $\mathscr{P}$ is given by [16]:

$$\textbf{(s,a)-rectangularity:}\ \ \mathscr{P} = \times_{(s,a)} \mathscr{P}_{s,a}, \quad \mathscr{P}_{s,a} = \{P_{s,a} : P \in \mathscr{P}\} \tag{3}$$

where $\mathscr{P}_{s,a}$ are called 'marginal' uncertainty sets. Essentially, under (s,a)-rectangularity, the uncertainty in transition probabilities at each state-action pair is independent. In some cases this assumption may prove to be too conservative [6, 16], leading us to consider other classes of uncertainty sets in Section 4. However, (s,a)-rectangular uncertainty sets may still be useful in many scenarios, especially due to their relative simplicity and some key theoretical results. E.g., for both finite and infinite horizon robust MDPs, an optimal policy that is both deterministic and stationary is guaranteed to exist [7, 10].

A very general algorithm for solving the robust policy improvement problem for (s,a)-rectangular uncertainty sets is presented in [6]. This robust value iteration algorithm proceeds in iterations given by

$$\mathbf{v}_s^{k+1} = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \lambda \min_{P_{s,a} \in \mathscr{P}_{s,a}} P_{s,a}^T \mathbf{v}^k \right\} \quad \forall s \in \mathcal{S} \tag{4}$$

where $\mathbf{v}_s^0 = 0$ for all $s \in \mathcal{S}$. (After the algorithm is done iterating, the optimal policy can be derived by taking the greedy policy.) In the remainder of this section, several specific examples of uncertainty sets that satisfy (s,a)-rectangularity are described (from Nilim & El Ghaoui 2005). Results and methods specific to each case are also discussed. Slightly different notation, matching that in Nilim & El Ghaoui 2005, will often be used in the remainder of this section.

3.1 Finite Scenario Models

We begin with finite scenario models, as described in Nilim & El Ghaoui 2005, which provide a very simple example of (s,a)-rectangular uncertainty sets. For each action $a \in \mathcal{A}$, define the uncertainty set

$$\mathscr{P}^a = \left\{ P^{a,1}, \dots, P^{a,L} \right\} \tag{5}$$

where $P^{a,k}$ are potential transition matricies. Letting $p_i^{a,k}$ be the i-th row of $P^{a,k}$, an equivalent formulation of this problem (meaning, the resulting optimal robust policy will be the same) is to write $\mathscr{P}^a$ as a product of convex hulls [10]:

$$\mathscr{P}^a = \bigotimes_{i=1}^{n} \mathbf{conv} \left\{ p_i^{a,1}, \dots, p_i^{a,L} \right\} \tag{6}$$

In this model, the minimization part of the robust value iteration algorithm (Equation 4) becomes simply $\min_{1 \leqslant k \leqslant L} \ v^T p_i^{a,k}$, which is very easy to solve; the complexity for each iteration of the algorithm is given as $O(SAL)$, where S and A are respectively the sizes of the state and action spaces [10].

3.2 Interval Matrix Models

Interval matrix model uncertainty sets are defined in the following form, where each p represents a row of the transition matrix [10]:

$$\mathscr{P} = \left\{p : \underline{p} \leq p \leq \overline{p}, p^T \mathbf{1} = 1\right\} \tag{7}$$

$\underline{p}$ and $\overline{p}$ are componentwise nonnegative n-vectors, which we assume satisfy $\underline{p} \leq \overline{p}$. The minimization in the robust value iteration algorithm (Equation 4) is also easily solved in this type of model, as we can simply write it as a linear program of the form

$$\sigma^* := \min_p v^T p : p^T \mathbf{1} = 1, \quad \underline{p} \leqslant p \leqslant \bar{p} \tag{8}$$

which could be solved using typical methods, e.g. by reformulating it as its dual.

3.3 Likelihood Model

Another way to model uncertainty sets is by the likelihood model. Let F^a be an empirical transition matrix (generated by observing the Markov chain transition out of each state many times under action a). One way to express uncertainty in this model was first introduced in Lehman et al. 1998. Denoting F^a and P^a as F and P for simplicity in this section [8, 10]:

$$\left\{P \in \mathbf{R}^{n \times n} : P \geqslant 0, P\mathbf{1} = \mathbf{1}, \sum_{i,j} F(i,j) \log P(i,j) \geqslant \beta\right\} \tag{9}$$

Here β represents the amount of uncertainty, which could be estimated through either resampling or a Gaussian approximation [10]. However, as written, the region in Equation 9 does not satisfy (s,a)-rectangularity. Nilim & El Ghaoui 2005 overcome this by projecting the likelihood region onto the rows of the transition matrix, creating a set which does satisfy (s,a)-rectangularity; as this is an overapproximation, however, the results produced will be more conservative. Because the log-likelihood function is separable, we can write the projection as [10]:

$$\mathscr{P}_i\left(\beta_i\right) := \left\{p \in \Delta^n : \sum_j f_i(j) \log p_i(j) \geqslant \beta_i\right\} \tag{10}$$

where

$$\beta_i := \beta - \sum_{k \neq i} \sum_j F(k,j) \log F(k,j) \tag{11}$$

The uncertainty set for P is then created by taking $\bigotimes_{i=1}^{n} \mathscr{P}_i\left(\beta_i\right)$ [10]. The minimization in the robust value iteration algorithm (Equation 4) can then be written in the following form,

$$\sigma^* := \min_p p^T v : p \in \Delta^n,\ \sum_j f(j) \log p(j) \geqslant \beta \tag{12}$$

and solved by considering the dual problem [10].

3.4 Entropy Models

The last example we describe in this section is the entropy model, in which the uncertainty set for p_i^a (the i-th row of P^a) is written as [10]:

$$\{p \in \Delta_n : D(p\|q) \leq \beta\} \tag{13}$$

where D gives the Kullback-Leibler divergence $(D(p\|q) := \sum_j p(j) \log \frac{p(j)}{q(j)})$ from some distribution q to p. Nilim & El Ghaoui 2005 assume that both q and β are positive so that the interior of the uncertainty set is nonempty. The β can be chosen using the methods described in Section 3.3 [10]. Similarly, the minimization in the robust value iteration equation can be solved by reformulating the problem as its dual. The result of this is the following optimality condition [10]:

$$\sum_j q(j) \exp\left(\frac{v(j) - \mu}{\lambda} - 1\right) = 1 \tag{14}$$

which produces the following optimal distribution [10]:

$$p^* = \frac{q(j) \exp(v(j)/\lambda)}{\sum_i q(i) \exp(v(i)/\lambda)} \tag{15}$$

The complexity analysis for the likelihood model and the entropy model are similar, and both are described in Nilim & El Ghaoui 2005.

4 Generalizations of (s,a)-rectangularity

So far, we have considered (s,a)-rectangular uncertainty sets, where the uncertainty between different state-action pairs is unrelated. As noted before, these uncertainty sets may be highly conservative and not reasonable in some scenarios. Now we consider two key generalizations of (s,a)-rectangularity: s-rectangular uncertainty sets [16] and factor matrix uncertainty sets [6]. Near the end of this section, we will also briefly discuss results pertaining to general uncertainty sets.

4.1 s-rectangularity

Recall that we defined $P_{s,a} = p(.|s,a)$ as the transition probabilities from each state-action pair. Under (s,a)-rectangularity, the uncertainty at each state-action pair was independent, with the uncertainty set given in terms of the Cartesian product of marginal uncertainty sets for each $P_{s,a}$ ($\mathscr{P} = \times_{(s,a)}\mathscr{P}_{s,a}$). One possible generalization of this would be to allow the uncertainty for each action in a certain state to be dependent (with the uncertainty at each state remaining independent). Wiesemann et al. 2013 introduce a class of uncertainty sets which satisfy this assumption, which they call ‘s-rectangularity’. If an uncertainty set $\mathscr{P}$ is s-rectangular, it can be written in the following form [16]:

$$\textbf{s-rectangularity: } \mathscr{P} = \times_s \mathscr{P}_s, \quad \mathscr{P}_s = \{(P_{s,1}, \ldots, P_{s,A}) : P \in \mathscr{P}\} \tag{16}$$

Note that since s-rectangularity is a generalization of (s,a)-rectangularity, any (s,a)-rectangular uncertainty set is also s-rectangular [16]. For a simple demonstration of an s-rectangular uncertainty set, we take an example used in Goyal & Grand-Clément 2018 and present it in Equation 17. Here, each $\mathscr{P}_s$ is described with a budget of uncertainty, with P_s^{nom} containing the nominal transition probabilities out of state s for each action a.

$$\mathscr{P}_s = \{P_s = P_s^{\text{nom}} + \Delta \mid \Delta \in \mathbb{R}^{A\times S},\ \|\Delta\|_1 \le \tau\sqrt{SA},\ \|\Delta\|_\infty \le \tau, P_s\mathbf{e}_S = \mathbf{e}_A,\ P_s \ge \mathbf{0}\} \tag{17}$$

While equation 16 is written in a very general form, Wiesemann et al. 2013 consider a more specific form for much of their results, where each $P_{s,a}$ is written in the following way:

$$p^\xi(.|s,a) = k_{s,a} + K_{s,a}\xi, \quad \text{where } \xi \in \Xi \subseteq \mathbb{R}^q,\ K_{s,a} \in \mathbb{R}^{S\times q},\ k_{s,a} \in \mathbb{R}^S \tag{18}$$

Using the form in Equation 18, the marginal uncertainty sets under (s,a)-rectangularity and s-rectangularity can be respectively written as:

$$\mathscr{P}_{s,a} = \{p^\xi(.|s,a) : \xi \in \Xi\}, \quad \mathscr{P}_s = \{(p^\xi(.|s,1), \ldots, p^\xi(.|s,A)) : \xi \in \Xi\} \tag{19}$$

Essentially, all the uncertainty in the marginals is represented by ξ, and Ξ is the uncertainty set for ξ. (Wiesemann et al. 2013 place a few important conditions on Ξ: its interior must be non-empty and it must “result from the finite intersection of closed half-spaces and ellipsoids”.)

Wiesemann et al. 2013 focus on the infinite horizon discoutned case, for which the worst case expected reward for a policy π can be written as $R(\pi)$ in Equation 2. An alternative way to write $R(\pi)$ while incorporating the form of Equation 18 is given in Equations 20-21. Wiesemann et al. 2013 write the value function v under a specific realization $\xi \in \Xi$ as

$$v_s(\pi;\xi) = \mathbb{E}^{p^\xi,\pi}\left[\sum_{t=0}^{\infty} \lambda^t r(s_t, a_t, s_{t+1}) | s_0 = s\right] \tag{20}$$

$R(\pi)$, the worst case expected total discounted reward, can then simply be written as

$$R(\pi) = \inf_{\xi\in\Xi}\{p_0^T v(\pi;\xi)\} \tag{21}$$

Wiesemann et al. 2013 consider two key problems: robust policy evaluation (how to calculate $R(\pi)$) and robust policy improvement (how to find a π^* where $R(\pi^*) = \sup_\pi R(\pi)$). To consider the policy evaluation problem, they begin by redefining their MDP as a 'Markov reward process' (essentially, a Markov chain with a reward at each state, where decision making is implicit rather than explicit [16]). The Markov reward process has modified rewards $\hat{r}$ and transitions $\hat{P}$ (see Wiesemann et al. 2013 for more details). This simplifies the notation somewhat, and allows them to produce the following result for the policy evaluation problem:

$$R(\pi) = \sup_{\vartheta:\Xi \xrightarrow{c} \mathbb{R}^S} \left\{ \inf_{\xi\in\Xi} \{p_0^T \vartheta(\xi)\} : \vartheta(\xi) \leq \hat{r}(\pi;\xi) + \lambda \hat{P}(\pi;\xi)\vartheta(\xi) \ \forall \xi \in \Xi \right\} \tag{22}$$

where $\vartheta : \Xi \xrightarrow{c} \mathbb{R}^S$ just means that ϑ is a continuous function from Ξ to $\mathbb{R}^S$. This is a difficult problem in the general case, but Wiesemann et al. 2013 show that under s-rectangularity, the optimal ϑ is a constant function. Furthermore, they show that its constant value can be found as the unique fixed point of the mapping ϕ, where

$$\phi_s(\pi; w) = \min_{\xi^s\in\Xi} \{\hat{r}_s(\pi;\xi^s) + \lambda \hat{P}_s^T(\pi;\xi^s)w\} \quad \forall s \in \mathcal{S} \tag{23}$$

Wiesemann et al. show that is possible to find this fixed point via robust value iteration, given by the algorithm $w^{i+1} = \phi(\pi; w^i)$. As with nonrobust value iteration, w^i will converge at a geometric rate to the unique fixed point [16]. A similar result holds for the finite horizon MDP problem, with a slightly different algorithm called 'robust backward induction' being used instead (see Wiesemann et al. 2013 for additional details).

The robust policy improvement problem can be treated similarly. Wiesemann et al. 2013 show that under s-rectangularity, the mapping φ in Equation 24 has a unique fixed point, which can again be calculated through robust value iteration.

$$\varphi_s(w) = \max_\pi \{\phi_s(\pi; w)\} \quad \forall s \in \mathcal{S} \tag{24}$$

Calling this fixed point w^*, the optimal policy is obtained by taking $\arg\max_\pi \phi_s(\pi; w^*)$ for each $s \in \mathcal{S}$ [16]. As before, Wiesemann et al. show that similar results extend to finite horizon MDPs, and provide complexity results for both scenarios. Wiesemann et al. also determine some key characteristics of the optimal policy π^* under s-rectangularity: they show that while a stationary optimal policy always exists, the optimal policy may be stochastic. This is in contrast to nonrobust MDPs and robust MDPs with (s,a)-rectangular uncertainty sets, where a deterministic, stationary optimal policy is guaranteed to exist [16].

4.2 Factor Matrix Models

While s-rectangularity should result in less conservative policies than (s,a)-rectangularity, there may still be a high level of conservatism since the uncertainty between different states is independent [6]. We now consider another generalization of (s,a)-rectangularity, which allows the uncertainty in transition probabilities to be related across different states: the factor matrix uncertainty set, as discussed in Goyal & Grand-Clément 2018.

In the factor matrix model, the transition probabilities from each state-action pair ($P_{s,a}$) can be written as a linear convex combination of r different factors $\mathbf{w}_i$ [6]:

$$P_{s,a} = \sum_{i=1}^{r} u_{sa}^{i} \mathbf{w}_i, \quad \text{where} \quad \sum_{i=1}^{r} u_{sa}^{i} = 1 \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A} \tag{25}$$

All the uncertainty on each $P_{s,a}$ is contained in the factors $\mathbf{w}_i$, while the u_{sa}^i are fixed (non-negative) coefficients for each state-action pair. More specifically, $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_r) \in \mathcal{W} \subseteq \mathbb{R}_+^{S \times r}$, where $\mathcal{W}$ is the uncertainty set for the factors. Since all $P_{s,a}$ depend on the same factors, the uncertainty on transition probabilities for each state-action pair is coupled. Goyal & Grand-Clément 2018 assume that $\mathcal{W}$ is convex and compact, and that the factors satisfy $\sum_{s'=1}^{S} w_{i,s'} = 1$ (each factor is a probability distribution over states).

Goyal & Grand-Clément 2018 note that factor matrix models can be used to represent any uncertainty set. As we will see later, the general case of uncertainty sets is intractable. To allow for tractability, Goyal & Grand-Clément 2018 place a restriction called r-rectangularity on the factor matrix uncertainty set $\mathcal{W}$:

$$\textbf{r-rectangularity: } \mathcal{W} = \mathcal{W}^1 \times \cdots \times \mathcal{W}^r \tag{26}$$

Essentially, this means that each factor $\mathbf{w}_i$ is chosen independently from a marginal uncertainty set $\mathcal{W}^i \subset \mathbb{R}_+^S$; each $\mathcal{W}^i$ is assumed to be convex and compact [6]. It should be noted that r-rectangularity is still a generalization of (s,a)-rectangularity, but is unrelated to s-rectangularity (there are s-rectangular uncertainty sets that are not r-rectangular, and vice versa [6]). Another difference between the two concerns the properties of the optimal policy: in the case of s-rectangularity, the optimal policy was stationary but not necessarily deterministic, whereas Goyal & Grand-Clément 2018 show that under r-rectangularity, a stationary and deterministic optimal policy exists.

Goyal & Grand-Clément 2018 provide an example of how to construct an r-rectangular factor matrix uncertainty set. They begin by decomposing the nominal transition probabilities P^{nom} into $\mathbf{W}^{\text{nom}}$ and $\mathbf{u}^{\text{nom}}$ by solving a nonnegative matrix factorization (NMF) program. They then create the uncertainty set on the factors by taking $\mathcal{W} = \mathcal{W}^1 \times \cdots \times \mathcal{W}^r$, where each $\mathcal{W}^i$ has a budget of uncertainty:

$$\mathcal{W}^i = \{\mathbf{w}_i = \mathbf{w}_i^{\text{nom}} + \delta \mid \delta \in \mathbb{R}^S, \ \|\delta\|_1 \leq \tau\sqrt{S}, \ \|\delta\|_\infty \leq \tau, \ \mathbf{e}_S^T \mathbf{w}_i = 1, \ \mathbf{w}_i \geq 0\} \tag{27}$$

Goyal & Grand-Clément 2018 also produce an efficient robust value iteration alogrithm for the robust policy improvement problem with factor matrix uncertainty sets. The algorithm proceeds in iterations given by

$$\mathbf{v}_s^{k+1} = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \lambda \sum_{i=1}^{r} u_{sa}^i \min_{\mathbf{w}_i \in \mathcal{W}^i} \mathbf{w}_i^T \mathbf{v}^k \right\} \quad \forall s \in \mathcal{S} \tag{28}$$

where the initial value function is set to be $\mathbf{v}^0 = 0$ ($\mathbf{v}_s^0 = 0 \ \forall s \in \mathcal{S}$). It should be noted that the one step reward $r(s,a)$, and the number of factors r, are unrelated. Complexity results for this algorithm are provided in Goyal & Grand-Clément 2018.

4.3 General Uncertainty Sets

Wiesemann et al. 2013 show that in the case of general uncertainty sets, the robust policy evaluation and robust policy improvement problems are strongly NP-hard for both finite and infinite horizon MDPs; the optimal policy is also no longer guaranteed to be Markovian. Although intractable to solve exactly (e.g., taking the approach in Section 4.1, there is no longer a guarantee that the optimal ϑ in Equation 22 is a constant function, as it was in the case of s-rectangularity [16]), Wiesemann et al. present some useful approximation methods for these problems. For instance, it is possible to get a lower bound on the worst case performance of a policy (Equation 22) by approximating a general uncertainty set $\mathscr{P}$ by an s-rectangular uncertainty set $\hat{\mathscr{P}}$ (specifically, let $\hat{\mathscr{P}}$ be the smallest s-rectangular uncertainty set containing $\mathscr{P}$), and then solving the problem for $\hat{\mathscr{P}}$ using robust value iteration [16]. Even better bounds could be obtained by solving Equation 22 over a larger class of functions than just constant functions, such as affine or piecewise affine functions [16].

Additionally, we point out that all the results so far have concerned convex uncertainty sets. As Wiesemann et al. 2013 note, considering nonconvex uncertainty sets produces very different complexity results. Even in the simple case of (s,a)-rectangularity, if $\mathscr{P}$ is nonconvex then the problem becomes strongly NP-hard [16]. The properties of the optimal policy may also change; e.g., for an infinite horizon MDP with a nonconvex s-rectangular uncertainty set, the optimal policy is no longer guaranteed to be stationary [16].

5 Extensions & Approximate Methods

Although we studied several key classes of uncertainty sets in this literature review, there are more that we did not cover. For example, Mannor et al. 2016 introduce k-rectangular uncertainty sets, which (like r-rectangularity for factor matrix models) allow for the uncertainty in transition probabilities to be related between different states while maintaining tractability. We also only touched upon methods for building uncertainty sets, with many of our prior examples of uncertainty sets based on simple deviations from predetermined nominal transition probabilities. More complex methods may have an additional computational burden, but may also result in smaller uncertainty sets and less conservative policies [11, 16], so we briefly mention a few such methods here. For example, Wiesemann et al. 2013 discuss how to generate uncertainty sets based on a series of observations of an MDP. Bayesian methods, such as those discussed in Petrik & Russel 2019, may also be useful for generating uncertainty sets.

Furthermore, most of the methods we have considered so far have been concerned with small, finite state and action spaces. As with nonrobust MDPs, extending robust MDPs to large or continuous state/action spaces usually requires the use of approximate optimization methods. There have been many recent advancements in the area of reinforcement learning for dealing with robust MDPs. For example, Tamar et al. 2013 introduce a robust approximate dynamic programming (RADP) method for dealing with large state space problems. Typical approximate dynamic programming methods make use of functional approximations so that quantities like the value function do not need to be calculated at every state; Tamar et al. 2013 take a similar approach to produce approximate solutions for robust MDPs.

Such approximate methods may be essential for dealing with large scale models of real world systems.

Although we have focused on uncertainty in state transitions so far, Wiesemann et al. 2013 note that it is also simple to consider uncertainty in the rewards (see Delage & Mannor 2010 for more details). Another reasonable place that uncertainty may appear is in the current state of the system, as in many applications it may be unrealistic to assume that a decision maker knows the exact state of the system when they are making decisions. In the POMDP (partially observed Markov decision process) framework, decisions must be made based on imprecise observations of the state of a system [3]. Along with considering state uncertainty, the POMDP framework also provides an alternative approach to considering uncertainty in state transitions [16]. Wiesemann et al. 2013 note that it is possible to consider uncertain model parameters (e.g., uncertain state transitions) by including this uncertainty in the states of a POMDP; however, they also note that this would likely lead to a very difficult and impractical optimization problem.

6 Empirical Work: Fisheries Management

6.1 Relevant Background and Nominal (Nonrobust) MDP

A key question in fisheries management concerns the optimal amount of fishing activity. Intensive fishing activity can result in large short-term harvests and profits, but may cause the fish population to be depleted quickly. Low intensity fishing is more likely to maintain the fish population or even allow it to recover, but at the cost of meeting current economic goals. We begin by describing a simple MDP that can model this scenario. This nominal MDP accounts for the stochastic transitions between states of the fishery at different levels of fishing activity, but assumes that the transitions are perfectly known. In reality, it is not known exactly how a certain level of fishing activity may affect a fishery. While the 'optimal' amount of fishing activity can be estimated from data, this data can be very limited and noisy. Thus, in Section 6.2 we describe a robust formulation of the MDP, which accounts for these uncertain transition probabilities. In Section 6.3, we compare the results obtained in the robust and nonrobust settings.

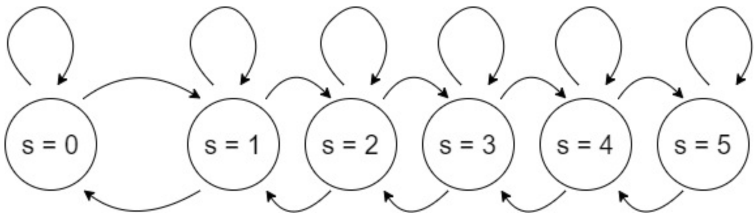

Figure 1: All states and possible transitions in our fisheries MDP

Figure 1 gives the basic framework of our MDP. Our MDP is not meant to represent any real fishery or fish species, but rather to provide a simple model of the trade-offs between fishing at different intensities. The chain proceeds in time steps representing years. The state s represents the current size of the fish stock (i.e., the population being fished). States $s = 1, \dots, 5$ represent stock sizes that are profitable to fish, ordered from small and nearly depleted ($s = 1$) to large and healthy ($s = 5$). On the other hand, $s = 0$ represents a stock that is 'commercially extinct': the stock is not gone entirely, but is so small that it is no longer profitable to fish. The action a represents the amount of fishing activity allowed by the fisheries manager during the year, with $a = 0$ representing no fishing activity, $a = 1$ representing low intensity fishing, $a = 2$ representing medium intensity fishing, and $a = 3$ representing high intensity fishing.

The (nominal) transition probabilities are given in Equations 29 and 30. While we selected these nominal transitions by hand, we believe they represent a range of realistic possibilities as the amount of fishing activity is varied. The transition probabilities reflect the highly stochastic nature of a fish stock: e.g., even under no fishing activity ($a = 0$), there is still a small chance the stock could decline, perhaps due to poor environmental conditions that year. Notably, we assume that due to the extremely small stock size, it is very hard to recover from the commercially extinct state $s = 0$: recovery is only possible under no fishing effort, and may take several years. Thus, it is very desirable to avoid this state whenever possible.

$$\mathbf{P}^{\text{nom}}_{\mathbf{s}} = \begin{array}{cccc} & p(s-1|s,a) & p(s|s,a) & p(s+1|s,a) \\ a=0 & .05 & .4 & .55 \\ a=1 & .1 & .45 & .45 \\ a=2 & .2 & .6 & .2 \\ a=3 & .6 & .35 & .05 \end{array} \quad \text{for } s = 1, \ldots, 4 \tag{29}$$

$$\mathbf{P}^{\text{nom}}_{\mathbf{0}} = \begin{array}{ccc} & p(0|0,a) & p(1|0,a) \\ a=0 & .8 & .2 \\ a=1 & 1 & 0 \\ a=2 & 1 & 0 \\ a=3 & 1 & 0 \end{array}, \quad \mathbf{P}^{\text{nom}}_{\mathbf{5}} = \begin{array}{ccc} & p(4|5,a) & p(5|5,a) \\ a=0 & .05 & .95 \\ a=1 & .1 & .9 \\ a=2 & .2 & .8 \\ a=3 & .6 & .4 \end{array} \tag{30}$$

The one-step reward $r(s, a)$ represents the total harvest or profit obtained during the year. For simplicity, we let the reward be $r(s, a) = sa$. This represents how more harvest/profit is possible when the stock size is larger, and when more fishing activity is allowed. Notably, the reward is zero under no fishing activity ($a = 0$), or when the stock is commercially extinct ($s = 0$). In the nominal setting, the manager's goal is to maximize the total expected discounted reward on an infinite time horizon. This means that in the nonrobust setting, we seek a (deterministic, stationary) policy π^* such that

$$v_s(\pi^*; \mathbf{P}^{\text{nom}}) = \sup_{\pi} v_s(\pi; \mathbf{P}^{\text{nom}}) \tag{31}$$

$$\text{where} \quad v_s(\pi; \mathbf{P}^{\text{nom}}) = \mathbb{E}^{\pi, \mathbf{P}^{\text{nom}}} \left[\sum_{t=0}^{\infty} \lambda^t r(s_t, a_t) | s_0 = s \right]$$

For simplicity we will assume that the initial state is always $s_0 = 5$, and we will use a discount factor of $\lambda = 0.9$. To solve the nominal MDP, we implement value iteration using the **MDPtoolbox** package in R [2]. The results of the nominal MDP will be discussed alongside the results of the robust MDP in Section 6.3.

6.2 Robust Formulation & Methods

As mentioned previously, it is not known exactly how a certain level of fishing intensity may impact a fishery. In the face of this uncertainty, fisheries managers may prefer to 'play it safe' to avoid worst-case fisheries outcomes (heavy depletion or extinction), even at the loss of some economic prospects. Fishers may also be risk-averse, preferring a smaller, guaranteed income with less risk over a larger expected income with greater risk. This motivates our robust formulation, in which case the manager's goal becomes to maximize the total expected discounted reward in the worst case scenario. More formally, we seek a policy π^* such that

$$R(\pi^*) = \sup_{\pi} R(\pi) \tag{32}$$

$$\text{where} \quad R(\pi) = \inf_{P \in \mathscr{P}} \{v_s(\pi; P)\} = \inf_{P \in \mathscr{P}} \mathbb{E}^{P,\pi} \left[\sum_{t=0}^{\infty} \lambda^t r(s_t, a_t) | s_0 = s \right]$$

and $\mathscr{P}$ is the uncertainty set. Due to limited time, we will focus on (s,a)-rectangular uncertainty sets. Recall that under this assumption, a deterministic and stationary optimal policy π^* is guaranteed to exist. We will assume that even under uncertainty, only transitions between neighboring states are possible: e.g., the probability of going from $s_t = 2$ to $s_{t+1} = 4$ will always equal 0, with no uncertainty. Thus, the uncertainty set on each $P_{s,a}$ will only concern the transitions $p(s-1|s,a), p(s|s,a),$ and $p(s+1|s,a)$. Define

$$\mathbf{P_{s,a}} = (p(s-1|s,a),\ p(s|s,a),\ p(s+1|s,a)) \in \mathbb{R}^3 \quad \text{for } s = 1, \ldots, 4 \tag{33}$$

$$\mathbf{P_{0,a}} = (p(0|0,a),\ p(1|0,a)) \in \mathbb{R}^2, \quad \mathbf{P_{5,a}} = (p(4|5,a),\ p(5|5,a)) \in \mathbb{R}^2 \tag{34}$$

Let the marginal uncertainty set for each state-action pair be $\mathscr{P}_{s,a}$. Similar to the examples provided in [6], we will look at $\mathscr{P}_{s,a}$ of the form (for $s = 1, \ldots, 4$),

$$\mathscr{P}_{s,a} = \{\mathbf{P_{s,a}} = \mathbf{P^{nom}_{s,a}} + \delta \mid \delta \in \mathbb{R}^3, \|\delta\|_\infty \le \tau, \|\delta\|_1 \le c\tau,\ \|\mathbf{P_{s,a}}\|_1 = 1, \mathbf{P_{s,a}} \ge \mathbf{0}\} \tag{35}$$

where the nominal transition probabilities are given in Equation 29. $\mathscr{P}_{0,a}$ and $\mathscr{P}_{5,a}$ are defined similarly, but with $\delta \in \mathbb{R}^2$ and the nominal transition probabilities given in Equation 30. The parameter τ represents the overall level of uncertainty, while the parameter c represents the budget of uncertainty. We will compare our results in the robust setting to the nominal setting, and look at how the results change as c and τ change.

We solve our robust MDP using robust value iteration for (s,a)-rectangular uncertainty sets, presented in Equation 4 and implemented in R. The discount factor ($\lambda = .9$) and starting state ($s_0 = 5$) are the same as in the nominal case. We found iterating the algorithm 100 times to be sufficient.

6.3 Results

Table 1 compares the optimal policy in the nominal (nonrobust) setting to the optimal policies of several robust scenarios. In the nominal case, the optimal policy involves increasing levels of fishing intensity as the stock size increases, with no fishing intensity when the stock is commercially extinct ($s = 0$) and high fishing intensity when the stock is large ($s = 5$). This pattern holds across all our robust policies. Fixing $c = 1$, under a small amount of uncertainty ($\tau = .3$) the optimal policy in the robust case matched the optimal nominal policy exactly. As the uncertainty increased further, the robust policies changed, but not in an obvious pattern. For example, at $\tau = .4$, the optimal policy was slightly less intensive than the nominal policy, while at $\tau = .5$, the optimal policy was more intense than the nominal policy. Under very high uncertainty (e.g., $\tau = 1$), the worst case for the fishery essentially became so bad that the optimal policy was to fish at the highest allowed intensity while it was still possible to fish at all.

Table 1: The optimal action to take at each state in the nominal (nonrobust) setting and several robust cases. Actions correspond to no (0), low (1), medium (2), or high (3) fishing intensity. Of the parameters defining the uncertainty set, τ was varied while $c = 1$ was fixed for all robust cases.

Case	$s = 0$	$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$
Nominal	0	0	1	1	2	3
$\tau = 0.3$	0	0	1	1	2	3
$\tau = 0.4$	0	0	1	1	1	3
$\tau = 0.5$	0	1	1	2	2	3
$\tau = 1$	0	3	3	3	3	3

In Figure 2, we examine how the value at our starting state ($s = 5$) in the robust setting compares to the corresponding value in the nominal setting. As expected, the value in the robust case decreases as c and τ increase (representing a greater amount of uncertainty). Considering the four robust cases in Table 1, the value function for each at the starting state was 68.9% ($\tau = .3, c = 1$), 58.9% ($\tau = .4, c = 1$), 52.5% ($\tau = .5, c = 1$), and 43.8% ($\tau = 1, c = 1$) of the corresponding value in the nonrobust setting. It is interesting to note that the robust optimal policy with $\tau = .3, c = 1$ was the same as the optimal nominal policy, even though the value in this case was reduced by almost a third compared to the nonrobust setting.

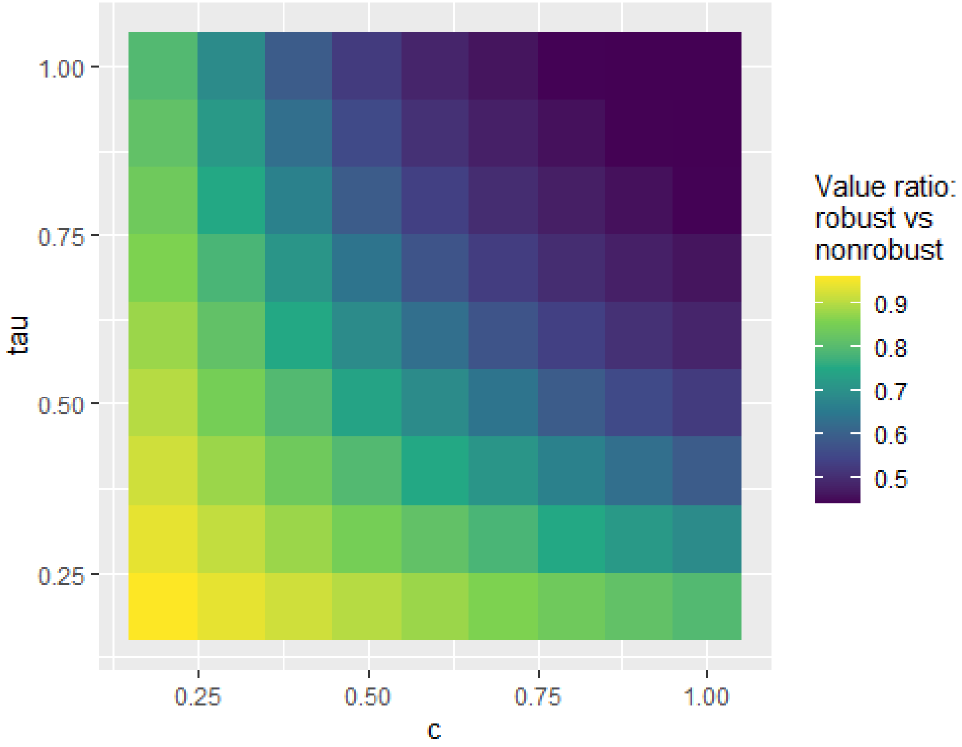

Figure 2: The value at the starting state ($s_0 = 5$) in the robust setting relative to the nominal setting. The parameters τ and c respectively define the overall level of uncertainty and the budget of uncertainty for the uncertainty set (see Equation 35).

7 Discussion & Conclusion

We discussed major contributions to the robust MDP literature in our review, notably focusing on three classes of uncertainty sets ((s,a)-rectangular, s-rectangular, and factor matrix models). Our review of (s,a)-rectangular uncertainty sets played an important role in our empirical work, where we explored a crucial question in fisheries management: what is the optimal level of fishing activity? Although the MDP we created to analyze this question was very simple, we were still able to provide interesting insights into this problem. Our results show that uncertainty in state transitions may greatly affect the optimal robust policy, and that the robust policy may call for the same, less, or more fishing intensity as compared to the optimal nominal policy (depending on the parameters defining the uncertainty set).

Our assumption of (s,a)-rectangularity was made to limit the computational burden and complexity of implementing the robust value iteration algorithm, but is likely too conservative for this situation. Wiesemann et al. 2013 state that uncertainty being dependent across actions may occur “when the actions of an MDP relate to varying degrees of intensity with which a task is executed”; this is certainly the case for our fisheries MDP, in which the action represents the intensity of fishing activity. A reasonable next step for this work would thus be to consider s-rectangular and/or factor matrix uncertainty sets (e.g., of the form in Equation 17 or 27), and compare those results to what we found under (s,a)-rectangularity.

Additionally, we would like to explore more complex fisheries models in future works.

Our empirical MDP was highly stylized to reduce the computational burden, and our nominal transition probabilities were not based on any real fishery dynamics. By basing the transition probabilities on a fisheries model that explicitly models populations growth, natural mortality, harvest, and so forth, we could better represent realistic transitions between states. Such an approach would require a much finer grid of states and actions to be useful, greatly increasing the computational burden of our work, but would also improve the specificity and realism of management advice given. Furthermore, such a model could be parameterized to provide management advice for a real fish species.

As we discussed in Section 5, there are additional sources of uncertainty we did not consider in our empirical work. Although not often considered by fisheries models, uncertainty in economic prospects (e.g., the price that fish will be sold at) could be represented by assuming that the rewards are uncertain. Furthermore, state uncertainty is a very relevant topic for fisheries management, as estimates of fish stock sizes are often based on limited data and modeling. Finally, beyond seeking the optimal level of fishing activity, there are many important fisheries modeling problems that could benefit from taking a robust approach as we have done here. Stock rebuilding, the placement of spatial closures, and the control of invasive species could all be optimized in a robust setting to explore risk-averse management strategies.

Code for the empirical work is available at: https://github.com/jeetjd/Robust-markov-decision-process-for-fisheries-management.git

References

[1] Bäuerle, N., Rieder, U. (2011). Markov decision processes with applications to finance. Springer Science & Business Media.

[2] Chades, I., Chapron, G., Cros, M., Garcia, F., & Sabbadin, R. (2017). MDPtoolbox: Markov Decision Processes Toolbox. R package version 4.0.3. https://CRAN.R-project.org/package=MDPtoolbox

[3] Chadès, I., Pascal, L. V., Nicol, S., Fletcher, C. S., & Ferrer-Mestres, J. (2021). A primer on partially observable Markov decision processes (POMDPs). Methods in Ecology and Evolution, 12(11), 2058-2072.

[4] Delage, E., & Mannor, S. (2010). Percentile optimization for Markov decision processes with parameter uncertainty. Operations research, 58(1), 203-213.

[5] Givan, R., Leach, S., & Dean, T. (1997). Bounded parameter Markov decision processes. In: Steel, S., Alami, R. (eds) Recent Advances in AI Planning. ECP 1997.

[6] Goyal, V., & Grand-Clément, J. (2018). Robust Markov decision process: Beyond rectangularity. arXiv preprint arXiv:1811.00215.

[7] Iyengar, G. (2005). Robust Dynamic Programming. Mathematics of Operations Research, 30(2), 257–280.

[8] Lehmann, E., & G. Casella. (1999) Theory of Point Estimation, Technometrics, 41:3, 274, DOI: 10.1080/00401706.1999.10485701

[9] Mannor, S., Mebel, O., & Xu, H. (2016). Robust MDPs with k-rectangular uncertainty. Mathematics of Operations Research, 41(4), 1484-1509.

[10] Nilim, A., & El Ghaoui, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. Operations Research, 53(5), 780-798.

[11] Petrik, M., & Russel, R. H. (2019). Beyond confidence regions: Tight Bayesian ambiguity sets for robust MDPs. Advances in Neural Information Processing Systems, 32.

[12] Shapiro, Alexander & Kleywegt, Anton (2002). Minimax analysis of stochastic problems, Optimization Methods and Software, 17:3, 523-542, DOI: 10.1080/1055678021000034008

[13] Tamar, A., Xu, H., & Mannor, S. (2013). Scaling up robust MDPs by reinforcement learning. arXiv preprint arXiv:1306.6189.

[14] White, D. J. (1985). Real applications of Markov decision processes. Interfaces, 15(6), 73-83.

[15] White, Chelsea C. & Eldeib, Hany K. (1994). ”Markov Decision Processes with Imprecise Transition Probabilities,” Operations Research, INFORMS, vol. 42(4), pages 739-749, August.

[16] Wiesemann, W., Kuhn, D., & Rustem, B. (2013). Robust Markov Decision Processes. Mathematics of Operations Research, 38(1), 153–183. http://www.jstor.org/stable/23358653